ACTA AGRARIA KAPOSVÁRIENSIS *(2020) Vol 24 No 1, 14-34.*
Kaposvári Egyetem, Agrár- és Környezettudományi Kar, Kaposvár
DOI: 10.31914/aak.2370

A genom összerakás elmélete és alkalmazása a gímszarvas genom projektben

Bana Ágnes Nóra *

Eötvös Loránd Tudományegyetem, Természettudományi Kar, Genetikai Tanszék, 1117, Budapest, Pázmány Péter sétány 1/C

***ABSTRACT* - Theory and application of genome assembly in red deer genome project**
Author: Ágnes Nóra Bana
Affiliation: Eötvös Loránd University, Faculty of Science, Department of Genetics, 1117 Budapest, Pázmány Péter sétány. 1/C

The entire inheritance of living organisms is contained in its genomic DNA, the sequence of which can be determined by various DNA sequencing methods. Researchers have long used clone-based methods for this purpose, but today the first, second (next-generation sequencing, NGS) and third-generation sequencing procedures are preferred. With the most widespread next-generation sequencing methods, a mammal's entire DNA can be read within hours. The single read sequences (reads) are only a few hundred basepairs, but summing up their lengths, they could cover the whole genome several times. The reads are assembled to longer sequences by using bioinformatics programs. The goal is to regain the sequence of the original DNA. First, contiguous sequences can be created from matching and overlapping reads, they are called contigs and contigs become settled for scaffolds. The sequences in the scaffolds are incomplete, there remain gaps, which are undetermined segments in the sequence. The most important farm animals have at least scaffold-level online genome assemblies. The most complete genome assemblies are ordered according to the pseudochromosomes. In the construction of pseudochromosomes, scaffolds are mapped to the genetic map of the organism in question, which was determined by genetic markers of known DNA sequences. In double-reference-guided assembling, *de novo* scaffolds are mapped/probed to the known genome of a closely related species and to the genetic map of the living entity targeted. The scaffolds are assorted and ordered following the colinearities and syntenies. This method is useful for completely new genomes. The annotation process describes the locations, structures, functions of the genes and other genomic structures along the reference genome resulted. The annotated reference genomes can be used in many ways, e.g. in animal breeding and husbandry, in the game management, in population genetics studies or in forensic cases/wildlife programs.

***Keywords*:** Genomics, Next-generation sequencing, Double-reference-guided genome assembling

BEVEZETÉS

A genom egy szervezet teljes örökítő információja, amelyet a DNS kódol (egyes vírusokban az RNS), a géneket és a nem kódoló szekvenciákat is magában foglalja. A genomika a genommal foglalkozó multidiszciplináris tudományág, amely a gének és a nem kódoló régiók kölcsönhatásait, a genomok szerkezetét, a gének elhelyezkedését vizsgálja és feltárja az egyes élőlények genomja közötti különbségeket. Az élőlények genomjában rejlő információk különböző

*CORRESPONDING AUTHOR
Eötvös Loránd Tudományegyetem, Genetikai Tanszék
✉ *1117 Budapest, Pázmány Péter sétány 1/C.,* ☎ *+36-1-372-2500/8073*
E-mail: bana.a.nora@gmail.com

bioinformatika eszközök segítségével válnak értelmezhetővé a kutatók számára.

A DNS szekvenálás során meghatározzuk egy DNS molekula nukleotid sorrendjét. Az új generációs szekvenáló laborok nagy mennyiségű szekvencia adatokat produkálnak, amelyek feldolgozásával a bioinformatika foglalkozik. A bioinformatika tárgykörébe tartozik a szekvencia illesztés, a statisztikai analízis (gén hosszúság, CG arány), a genom annotáció (ORF, gén predikció, promóter analízis), szekvencia adatbázisok használata, szekvencia keresés, makromolekulák háromdimenziós modellezése és a fehérjék közötti kapcsolatok feltárása. A Next-generation sequencing (NGS, Új-generációs szekvenálás módszereknek köszönhetően napjainkban elegendő csupán nyolc munkaóra és körülbelül 1000-1500 USA $ egy emlős teljes genomjának a szekvenálásához. A leolvasott és összerakott szekvenciákat a bioinformatikusok igyekeznek nemzetközileg elismert, online elérhető adatbázisokba feltölteni. Az egyik ilyen jelentős publikációs és szekvencia adatbázis az NCBI (National Center for Biotechnology Information), ahol a feltöltött genomok száma folyamatos növekedést mutat.

2018. január 23.-án a svájci Davos városában tartott találkozón a Világgazdasági Fórum több fontos genomikai intézettel lépett szövetségre. A konferencián megfogalmazták a negyedik ipari forradalmat megalapozó Föld BioGenome Project (EBP) célkitűzéseit, amely lényegében a világ minden eukarióta élőlényének (összes növény, állat és egysejtű szervezet) a teljes genom szekvenálása és egy fenntartható biogazdaság létrehozása. Ezek az intézkedések fontos szerepet játszhatnak mintegy 20000 veszélyeztetett faj megmentésében. A projekt előreláthatóan 10 évet és 4,7 milliárd dollárt fog igénybe venni (*Lewin és mtsai.,* 2018).

Régen kimagasló eredménynek számított egy-egy gén teljes szekvenciájának leírása, manapság azonban az élőlények teljes genomjának ismertetésére úgynevezett WGS (Whole Genome Sequencing) projektek megvalósítására törekszenek a kutatók világszerte. Napjainkban a genetikai modell állatokon kívül, több mezőgazdaságban kiemelkedő növény és állatfaj teljes genom szekvenciája is elkészült például házi sertés, szarvasmarha, juh, rizs, búza, kukorica stb.

Magyarországon a gímszarvas genomprogram keretében megvalósult Magyarország első igazi, nemzetközileg elismert genomprogramja, amelynek kapcsán létrehoztuk a világ első szarvas genom programját (CerEla1.0), amely kromoszómákba rendezve, valamint a centromeron pozíciókhoz orientálva készítettünk el (*Bana és mtsai.*, 2018).

TELJES REFERENCIA GENOM ÖSSZEÁLLÍTÁS MENETE

Laboratóriumi munka

A bioinformatikai analíziseket szinte minden esetben megelőzi a laboratóriumi munka. A teljes genomi DNS kinyerése és a mintavételezés meg kell feleljen a humán etikai vonatkozásoknak, vagy az állat jóléti törvényeknek. A minták tárolása és a DNS izolálási protokoltól és a későbbi felhasználástól függ. Az izolálás eredményeképpen kapott DNS tisztaságát és mennyiségét nanodrop fotométeren ellenőrzik le. A gímszarvas genomprogram esetében a DNS mintát a Kaposvári Egyetem Vadgazdálkodási Tájközpont Bőszénfai Szarvas farmján, természet közeli körülmények között élő, 7 éves, kapitális gímszarvasbika (fülszáma: Crot. N.o. 3016) 10 ml vére szolgáltatta, amelyet a mintavétel után EDTA pufferben tároltunk.

DNS könyvtárak

A vírusok és baktériumok szekvenálása laboratóriumi, klón alapú módszerrel kezdődött el. A klón alapú, hierarchikus szekvenálás során a genomi DNS-t restrikciós enzimekkel vagy különböző mechanikai eljárásokkal véletlenszerűen apró, 40-150 kbp-os, átfedő darabokra tördelik szét. Ezek a darabok nagyon gyakran STS-eket (Sequence-Tagged Sites), az adott genomi pozícióra jellemző rövid, egyedi szekvencia részeket is tartalmaznak. A kis DNS fragmentumokat plazmid vektorba klónozzák be és a recipiens baktérium sejtekben, általában *Escherichia coli*-ban szaporítják fel. Ezt az eljárást nevezzük molekuláris klónozásnak, az így létrejött a faj teljes genomját tartalmazó BAC (Bacterial Artificial Chromosome) gyűjteményt pedig BAC könyvtáraknak hívják. Az egyedi BAC DNS darabokból további tördeléssel még kisebb (szub) fragmentumokat hoznak létre, és ezek újabb felszaporítással szubklón könyvtárakat konstruálnak. A fragmentumok első 500 bp-ját leolvassák (read szekvencia). A hosszabb szekvencia részeket, a contigokat és a szupercontigokat az átfedő DNS-régiók alapján elektronikus úton állítják össze például PHRAP szoftver segítségével (*de la Bastide és McCombie*, 2007). A contig szó összefüggő, ismert bázissorrendű, hosszabb nukleotid sorozatot jelöl. A *Haemophilus infuenzae* baktérium teljes genomjának a feltárása „shotgun" eljárással ment végbe (*Fleischmann és mtsai.*, 1995). Az egész genom véletlenszerű „shotgun" feldarabolása úgy történik, hogy a DNS-t átpasszírozzák egy vékony csövön és az így kapott 2-10 kbp -os a darabokat szekvenálják meg a végeiről indulva a közepe felé. Emiatt a leolvasott DNS darabkák többszörösen fedik le a teljes

genomot, ami a contigok közötti hézagok számának nagymértékű csökkenéséhez vezet. A contigok közötti részeket PCR segítségével határozzák meg. A WGS (Whole genome shotgun) contigok számítógépes összeszereléséhez (assembly készítés) legelőször TIGR Assembler számítógépes szoftvert alkalmaztak (*Sutton és mtsai.*, 1995).

Első, második és harmadik generációs szekvenálás

Az 1970-es évek végén váltak elérhetővé az úgynevezett első generációs szekvenálási technikák. Ezen klasszikus módszerek közé sorolható a Sanger és Coulson nevével fémjelzett +/- szekvenálás amellyel, először olvasták le egy vírus, a phi X174 bakteriofág teljes genomját (*Sanger és Coulson*, 1975).

Hagyományos kémiai módszereket használ fel a Maxam- és Gilbert-féle kémiai hasítás (*Maxam és Gilbert*, 1977). Ilyenkor a DNS cukor-foszfát gerince mentén történik meg a specifikus hasítás (A+G, G, C+T, C bázisok után). Az eltérő méretű méretű DNS darabokat denaturáló poliakrilamid gélelektroforézissel választják el és autoradiográfiával detektálják. A végeredmény a „DNS fragmentek létrája", amelyből leolvasható a szekvencia.

A legismertebb első generációs eljárás a Sanger-féle láncterminációs szekvenálás (*Sanger és mtsai.*, 1977). Az egyszálúsított templát DNS-hez szekvenáló primereket hibridizálnak, majd DNS polimerázt, azonos mennyiségű négyféle dezoxinukleotid-trifoszfátot (dNTP) és négyféle fluoreszcens festékkel (fluoroflórral) jelölt didezoxinukleotid-trifoszfát (ddNTP) molekulát kevernek a PCR reakció közegbe. A DNS polimerizáció alatt, vagy a komplementer dNTP vagy ddNTP épülhet be az új szálba, a nukleotidok koncentrációjának arányától függően. Amikor dNTP illeszt̋ődik be a szálba a szintézis tovább folytatódik, mert a dNTP 3' végén -H csoport található. Amikor ddNTP kerül be akkor a dNTP 3' végén lévő reaktív -OH csoport miatt a szintézis leáll. Ennek a folyamatnak a neve a láncterminációs reakció. Az új DNS szálak szétválasztására kapilláris gél elektroforézist használnak (*Dovichi és Zhang*, 2000). Az automata fluoreszcens szekvenálás nagy előnye, hogy a DNS amplifikálása PCR eljárással valósítható meg.

A második generációs vagy új generációs szekvenálási (next-generation sequencing, „NGS") technológiáknál a DNS amplifikációja mindig valamilyen PCR reakcióval történik meg, és az utolsó (szekvenáló) PCR során leolvasott szekvenciákat, a readeket nagyteljesítményű számítógépek kezelik és dolgozzák fel. Az NGS-nél több millió szekvenálási reakció zajlik egyidőben, ami gyorsabbá és olcsóbbá teszi ezt az eljárást. Egy emlős teljes genom szekvenálása csupán pár órát vesz igénybe és napjainkban körülbelül 300.000 Ft-ba kerül.

A leolvasott readek hossza 50-700 nukleotid között van. Amennyiben csak a DNS fragmentumok egyik végét szekvenálják meg, úgy single end sequencingről beszélünk. Paired-end szekvenáláskor viszont a DNS szekvenciák mindkét végét meghatározzák. A szekvencia leolvasása nagyfelbontású CCD kamera segítségével történik. Az NGS technológiákról általában elmondható, hogy nagy áteresztőképességű (high-throughput, „HTP") módszerek is egyben, hiszen több minta leolvasása párhuzamosan, egyidőben zajlik le. A legismertebb második generációs szekvenálások közé tartoznak az Illumina (Solexa) sequencing, a SOLiD (Sequencing by Oligonucleotid Ligation and Detection) sequencing, a Pyrosequencing (454) és az Ion Torrent sequencing. Az Illuminánál a DNS lánc épülése szintézissel zajlik és DNS szálak úgynevezett hídamplifikációval sokszorozódnak fel. Az eltört DNS szálak egy szilárd szekvenáló lemezhez rögzített oligonukleotidhoz kapcsolódnak hozzá, majd a polimerizáció során a DNS templátról átírt szálak a denaturációs szakasz után lehajlanak egy másik rögzített primerhez, azzal hibridizálnak és az elongációs folyamat révén duplikálódnak. A keletkező PCR termékek „klaszterekbe" rendeződnek. A feldúsult DNS fragmentumok leolvasása szekvenáló primerek és eltérő színű fluorescens festékkel jelölt dNTP-kel történik meg. Amikor beépül egy nukleotid annak fluorescens villanását észleli és rögzíti a kamera. A paired-end szekvenálás egy specifikus típusa a mate pair párok létrehozása, amelynél a DNS-t először 2-5 Kbp nagyságú darabokra törik, végüket biotinilálják majd a DNS szálakat cirkularizálják és végül még kisebbre fragmentálják. Az új fragmentumok közül néhány tartalmazza mindkét biotinilált mate pair szegmenst. A továbbiakban a paired-end szekvenálásra jellemző metódika alapján járnak el (*Bentley és mtsai.*, 2008). A gímszarvas genom program készítésekor 4 paired-end és 2 mate pair nyers read szekvencia könyvtárat hoztunk létre. A SOLiD és Pyroszekvenálás közös eleme, hogy az „emulziós" PCR reakció során a templátokat szilárdfázisú polisztirén gyöngyökhöz kötik hozzá. A SOLiD eljárásnal a szekvenálás az oligonukleotid próbák ligálásával történik meg, míg a Pyroszekvenáláskor DNS szintézis segítségével. Ez utóbbinál dNTP-khez kötött pirofoszfát a nukleotid szálba épülésekor leválik és a szulfuriláz enzimhez kapcsolódik, amely ATP-t állít elő. Az ATP-ből pedig luciferin jelenlétében oxiluciferint jön létre és ez a reakció fényvillanással jár (kemilumineszcencia)

(*Pevsner*, 2015). Az Iontorrent esetében a nukleotid DNS szálba épülése hidrogénion kilépéssel jár, amelyet félvezető chip-en sok, apró pH mérő érzékel, amit egy számítógépnek továbbít így kémiai jel alakul át digitális jellé (*Rusk*, 2011).

A harmadik generációs nanoporus szekvenálás valós időben (real-time) megy végbe, önálló DNS darabokon. Emiatt nincs szükség PCR amplifikálásra.

A DNS láncot elektroforézisel 1 nanométer átmérőjű membránpóruson húzzák át, ami bázisonként változtatja meg a membrán elektromos potenciálját. A változást mutató jelet nagy érzékenységű detektor észleli és küldi tovább a szekvenátor készülékhez (*Niedringhaus és mtsai.*, 2011).

A szekvenálás eredményeként a read szekvenciákat és a rájuk jellemző tulajdonságokat tartalmazó fastq kiterjesztésű fájlok jönnek létre. A leolvasott readek minőségét például a FastQC programmal ellenőrizhetjük le (*Andrews*, 2010). A szekvenáláskor leolvasott nukleotidok minőségi mutatója a Phred score, ami egy logaritmusos érték. Azt mutatja meg, hogy mekkora az esélye annak, hogy egy bázist helytelenül határoztuk meg. A trimming során a rossz minőségű readeket eldobjuk, mivel zavarhatják a genom analízis további lépéseit.

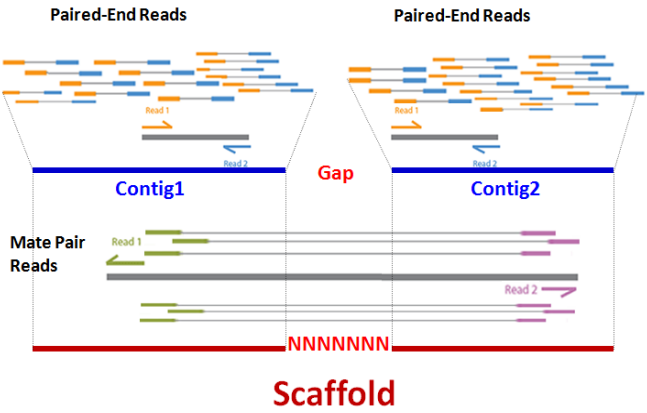
Contig és Scaffold generálás

A legtöbb új-generációs szekvenálás során kapott read méret 100-150 bp körüli, viszont a későbbi genom analízisekhez nélkülözhetetlen a hosszabb, összefüggő DNS szekvenciák elemzése. Amennyiben már ismert referencia genomú faj egyedéből származott a DNS minta, abban az esetben reszekvenálásról beszélünk. A referencia genomot több szekvencia közös, összeilleszkedő részei alapján *in silico* generálták. A referencia genom mindig haploid genom, amely a jelentősebb bioinformatikai, genomikai portálokról elérhető és letölthető. A re-szekvenálásnál a readeket például bwa-mem programmal illeszthetjük fel a már létező referenciához (*Li*, 2013). Ilyenkor azonban csak a kisebb eltérések, például a mikroszatelliták (egymás után elhelyezkedő, 1-6 bázispár hosszúságú ismétlődésekből álló DNS szekvenciák), indelek (néhány bázispárnyi inszerciók vagy deléciók) és az SNP-ek (Single Nucleotide Polymorphism, vagyis egyetlen bázispárt érintő nukleotid-polimorfizmus) lokalizálásának van értelme. Ugyanis az illesztő programok nem veszik figyelembe a nagyobb genomi átrendeződéseket. Éppen ezért sem célszerű a referencia genom nélküli fajoknál sem, hogy a readeket egyszerűen csak egy rokon faj genomjához illesszük fel, hiszen nagyon sokszor még kromoszóma szám tekintetében sem egyezik meg a két faj egymással. A jó minőségű, nagy átlagos lefedettségű readek önmagukban is elegendők hosszabb összefüggő szekvenciák úgynevezett contigok „*de novo*" létrehozására. A DNS-szekvenálás lefedettsége (vagy mélysége) azoknak az egyedi readeknek a számát jelöli, amelyek egy adott nukleotidot tartalmaznak, és átfednek a rekonstruált szekvenciában.

Az assembler szoftverek egy része „mohó vagy greedy" algoritmus alapján működik. Az algoritmus minden lehetséges read páronkénti illesztést figyelembe vesz a szekvencia egyezés alapján. Ezután a két legjobban átfedő readet kiválasztja és összevonja (a művelet bioinformatikai elnevezése a „merge"). Ezt a lépést addig ismétli, amíg el nem fogynak a readek. Ez a megoldás rendkívül gyors, ámde a tévedéseit nem korrigálja. A többi illesztő program a De Bruijn gráf elvet használja, vagyis a readeket kisebb, „k" nukleotid hosszúságú részekre bontja (ezek lesznek a k-merek, amelyek a gráf csúcsait adják) és két csúcsot akkor köt össze egy irányított éllel, ha az első csúcs utolsó k-1 betűje megegyezik a második csúcs első k-1 betűjével. Ezek a k-merek hosszabb szekvenciákká fűzik össze a readeket. Problémát okozhatnak az illesztő programok számára az ismétlődő DNS szakaszok, és a rossz helyre, illetve rossz orientációban elhelyezett readek. Néhány fontosabb assembler program: SPAdes, Ray, ABySS, Trinity, HGAP, Falcon, Canu, Hinge, Assemblathon, Racon, Celera WGA, Edna, Euler, MIRA, Newbler, SOAPdenovo, ALLPATHS-LG, Discover. A programok nevét a google keresőbe beírva részletes leírásokat találunk.

A gímszarvas genom program esetében az ALLPATHS-LG illesztő programot használtuk, amelyet Illumina szekvenálásból származó 100 bp-os vagy annál nagyobb hosszúságú readek kezelésére fejlesztettek ki (*Gnerre és mtsai.*, 2011). Működési feltétele, hogy a szekvenálásból legalább két, páros read könyvtár álljon rendelkezésre, amelyekből az egyik Paired-end, a másik Mate pair legyen, és mindkét könyvtárra minimum 45-szörös átlagos lefedettség legyen jellemző. Eredményként megkapjuk az assemblyt (efasta és fasta kiterjesztéssel), illetve egy riport fájt („ALLPATH report"), ami a legfontosabb statisztikákat írja le. A riportban az assembly minőségére vonatkozó információkat is láthatjuk, ilyen például az N50-es mérőszám contigokra vagy scaffoldokra vonatkoztatott értéke, ami egy súlyozott medián érték. Azt a scaffold hosszúságot adja meg, aminél hosszabb scaffoldok az összes assembly felét teszik ki.

Maga az assembly tehát állhat contigokból és scaffoldokból is. A contigok szinte nukleotid bázisokból felépülő hosszú szekvenciák, amelyek a readek átfedései révén jönnek létre. Azonban ahol a program nem talált átfedő readeket, ott az egymástól 2-5 Kbp távolságra található mate pair read párok tagjait hívja segítségül, amelyek szekvencia vázakká fűzik fel a contigokat. Az így összefűzött contigokból kialakuló szekvencia elemeket scaffoldoknak nevezzük. A contigok közötti ismeretlen bázisokat általában N-nel jelöljük, és gap régióknak nevezzük. A scaffoldok tehát mate pair read párokkal felfűzött, contigokból állnak (1. *ábra*).



1. ábra
A contig, scaffold, assembly *de novo* létrehozásának sematikus rajza.
Figure 1. Schematic drawing of the de novo.creation of contig, scaffold, assembly.

Genom annotáció

A szekvenált, contigokká és scaffoldokká összerakott és leolvasott DNS lánc hosszú nukleotidbázis betűsorozata önmagában nem informatív, ezért van szükség például a kódoló régiók megkeresésére és elnevezésére. A gének elnevezése a múltban empirikus módon történt, ami ahhoz vezetett, hogy egy gén többféle alternatív nevet is kaphatott. Emiatt a név nagyon sokszor nem specifikus, kivéve a humán géneket, ahol a HUGO (Human Genome Organisation) gene nomenclatura committee igyekezett egységesíteni a különböző elnevezéseket. Az annotált, kódoló régiók szekvenciái többnyire elérhetők olyan online genom adatbázisokból, mint például az NCBI, az Ensembl, Mouse Genome Informatics, FlyBase, és a WormBase. Ezekben az adatbázisokban minden szekvenciához tartozik egy egyedi azonosító szám, ami csupán egy entitást határoz meg, Minden adatbázisnak megvan a maga szekvencia azonosító formátuma, ami betűk és számok kombinációjából áll. A különböző adatbázisok gyakran átjárhatók, kereszthivatkozhatók egymással.

A gének azonosításakor, az annotálási eljárás során megkeresik a genomban a releváns DNS szekvenciákat és különböző biológiai információkat rendelnek hozzájuk. A folyamat három fő lépésre osztható:

1. A nem fehérje kódoló régiók azonosítása a genomban

2. Gén predikció, vagyis a lényeges genomikai elemek identifikációja

3. Biológiai információk csatolása ezen genomikai elemekhez (*Stein*, 2001).

A biológiai információ szempontjából megkülönböztetjük egymástól strukturális és funkcionális jellegű annotációt. Strukturális adatnak számít a kromoszómán való lokalizáció, a pontos génszerkezet, vagyis az exon-intron határok, UTR, promóter régiók megkeresése és a szabályozó régiók leírása. A funkcionális annotáció kapcsán meghatározzák az adott gének vagy DNS szakaszok biokémiai, különböző biológiai, regulációs, expressziós és interakciós tulajdonságait. A folyamatban egyaránt használhatnak kísérletes adatokat és számítógépes vizsgálatokat is.

Az automatikus annotációnál ezeket a lépéseket kizárólag számítógépes elemzéssel végzik el ellentétben az emberi szakértelemmel járó kézi annotációval. Ideális esetben ezeket a módszereket együttesen alkalmazzák. Az automatikus annotálásnál gyakran használnak genetikai homológiát kereső eszközöket, például a BLAST illesztőprogramcsomagot. Ilyenkor a különböző adatbázisokból és akár más fajokból is származó gén, cDNS, EST, RNA-seq, protein, mRNS szekvenciákat illesztik fel a teljes genomra, ezután a különböző szekvencia találatok átfedő részeiből klasztereket képeznek, majd meghatározzák ezek pontos genomi helyzetét. Mivel ezekben az esetekben valós kereső szekvenciák segítségével találják meg a kódoló részeket a genomban, ezért ezt a módszert bizonyíték alapú gén predikciónak nevezzük. Ezzel ellentétben az ab initio gén predikciónál nincs szükség külső kereső szekvenciákra, az annotáló program például a promóter régiókra, transzkripciós starthelyekre, exon-intron határokra, poliadenilációs helyekre utaló speciális szekvencia jelek és jellemző statisztikus tulajdonságaik alapján jelöli ki ezek helyét az ismeretlen genomban. A legtöbb nemzetközi online genomikai adatbázis saját annotációs projektekkel és pipeline-nal rendelkezik, ilyen például DNS-elemek enciklopédia (ENCODE), Entrez Gene, Ensembl, GENCODE, Gén ontológiai konzorcium, GeneRIF, RefSeq, Uniprot. A gén predikciós eszközökből is elég sok áll rendelkezésre, amelyekből az adott genomnak megfelelőt érdemes kiválasztani, íme néhány program: FusionSeq , GAAP, GENSCAN, GENEID, GENEMARK, JIGSAW, Artemis, AUGUSTUS, EuGene, MAKER, MAKER2, OmicX, PseudoPipe, TAIR.

Kromoszómába rendezés

Egy faj teljes genom szekvenálásából elsőként elkészült scaffold és contig DNS szekvenciák összesége vagyis a *de novo* assembly csak korlátozott mértékben használható genomikai analízisekre, mivel a gének működését befolyásoló elemek egymástól egészen nagy távolságokra helyezkedhetnek el a DNS láncon. A legmegbízhatóbb eredményeket a lehető leghosszabb úgynevezett szuper scaffold létrehozásával és vizsgálatával érhetjük el, ami nem más, mint a faj egy kromoszómájának teljes nukleotid sorrendje. Sok élőlény DNS szekvencia összerakása csupán a scaffold assembly szintig jut el, mert nem áll rendelkezésre megfelelő géntérkép. A géntérképek morgan vagy centimorgan egységben adják meg a gének vagy a markerek, azaz ismert DNS szekvenciák egymástól való távolságát és kromoszómális elhelyezkedését. 1 Morgan (M), annak a két pontnak/génnek a távolsága, amelyek között 1, a crossing-overek átlagos gyakorisága, 1 centimorgan (cM) távolság pedig 1% átlagos crossing-over gyakoriságnak felel meg. A genetikai térképeken 1 cM egy genetikai térképegységet jelent. Géntérkép hiányában, ha nagyobb szekvencia egységekre van szükség, megoldást jelenthet egy közel rokon faj teljes referencia genom szekvenciája. A tyúkidomú nyírfajd genom összeszerelésénél a generált scaffoldokat a csirke referencia 28 autoszómájára és Z kromoszómájára illesztették, hogy fel tudják tárni a homológ gének közötti összefüggéseket (*Wang és mtsai.*, 2014). A géntérkép nélküli kromoszómákba rendezéshez szerencsés esetben is csak bonyolult citológiai eljárásokkal juthatunk (jelölt scaffoldok in situ hibridizálása kromoszóma preparátumhoz, multicolor in situ hibridizáció).

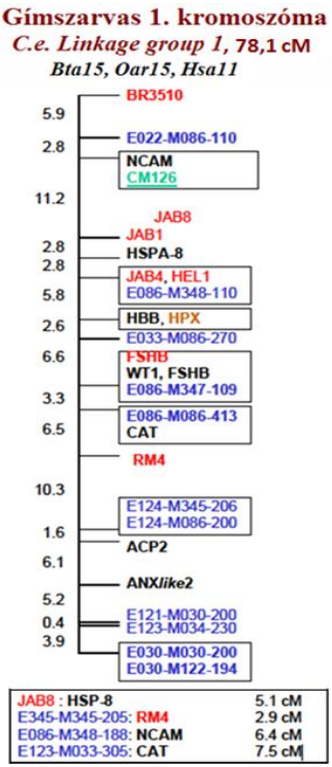
Amikor azonban nem elérhető valamely rokon faj genomja elméletileg átfedő BAC vagy shotgun klónokból fel lehet építeni egy teljes genomot, de a valóságban mindig szükség volt citológiai referenciákra, például sávozott kromoszóma preparátumokra és in situ hibridizálásokra, géntérképekre.

Szerencsésebb helyzetekben az adott faj rendelkezik saját géntérképpel, amelyek markereivel könnyen kihalászhatók és térképre illeszthetők a marker szekvenciát tartalmazó *de novo* scaffoldok (*Shulaev és mtsai.*, 2011). Értelemszerűen ilyenkor minél sűrűbb géntérképre, és minél nagyobb scaffoldokra van szükség.

A gímszarvas esetében lehetőség nyílt a kettős referencia vezérelt genom összerakásra, mivel egyaránt rendelkezésre állt egy közel rokon faj, a szarvasmarha jól annotált referencia genomja és egy rekombináns gímszarvas géntérkép is. A *de novo* létrehozott scaffoldok szavatolták, hogy a gímszarvasra jellemző sajátos szekvencia részek megmaradjanak. E scaffoldok gímszarvas géntérképre és szarvasmarha genomra illesztése során pedig megkaptuk a

scaffoldok és a rajtuk elhelyezkedő fontos genetikai struktúrák (gének) egymás utáni sorrendjét. A közel rokon fajok közül azért a szarvasmarhát választottuk, mert jól annotált, és a szarvasmarha genom konzervatív régiói 92%-98% hasonlóságot mutatnak a gímszarvas cDNS-sel (*Gyurján és mtsai.*, 2007, *Stéger és mtsai.*, 2010). A gímszarvas géntérkép annyiban is különleges, hogy nem használhattak háziasított gímszarvas fajtákat a létrehozására, helyette az un. „interspecifikus back-cross" módszert alkalmazták, mivel a szarvasfélék közeli fajai keresztezhetők egymással, és fertilis fajhibrideket hoznak létre. Ezt láthatjuk a milu (*Elaphurus davidianus*) és a gímszarvas (*Cervus elaphus*) interspecifikus keresztezésénél, ahol a Haldane szabállyal (Az emlős fajhibridek hímjei, XY, a „heterogamétás szex", általában sterilek, a női egyedek XX, a „homogamétás szex" ugyanakkor fertilisek) ellentétben nemcsak a nőivarú, hanem a hím és női ivadékok egyaránt termékenyek. Továbbá a genomjukban nagyszámú fajspecifikus genetikai variáció található, ami jellemezi az adott lókuszt. E markerek allélikus variációi jól kimutathatók DNS diagnosztikai eljárásokkal, ha az allélok hossza különbözik (például az elektroforetikus vándorlási sebesség alapján, azaz PAGE vizsgálattokkal) és felhasználhatók a rekombinációs gyakoriságok meghatározásához és a géntérképezéshez. A térképező populáció létrehozása két új-zélandi szarvasfarmon történt, ahol az F1 nemzedék milu és gímszarvas hibrid szarvasbikákat kereszteztek gímszarvas ünővel, amelynek eredményeképpen 351 back-cross utód született. A mesterséges megtermékenyítésből született (F2) szarvasborjak adták a térképező populációt a géntérkép elkészítéséhez. Meghatározták a DNS marker variációk között a rekombinációk gyakoriságait, ebből cM távolságokat számoltak és így tudták megszerkeszteni a gímszarvas autoszómák géntérképét. A rekombináns kromoszómák a meiózis profázisának I szakaszában jöhetnek létre oly módon, hogy a homológ kromoszómapár karjai között átkereszteződés (crossing-over) és kromoszómális szakasz csere történik („törés-újraegyesülés modell"). Így tehát új allélkombinációjú (R, rekombináns) kromoszómák keletkeznek a parentális (P) azaz nem rekombináns, eredeti allékombinációt hordozó kromoszómák mellett. A rekombinációs gyakoriságot (r) úgy kapjuk meg, hogy a rekombináns gaméták összességét osztjuk az összes gaméta számával, azaz r= R/P+R. A rekombinációs gyakoriság (r) kimutatása irányított keresztezésekkel történhet, ahol az utódokból következtethetünk vissza a parentális és rekombináns gaméták gyakoriságára (tesztelő keresztezés, „teszt-cross" segítségével). A gének és markerek a genetikai távolságon kívül, egyéb tulajdonságaikban, például a sorrendjükben is összefüggést mutatnak a rekombinációs gyakoriságokon alapuló kapcsoltsági térképeken és a

szekvenálás eredményeképpen összeállított fizikai térképeken, vagyis a géntérképi pontok és kromoszómán található lókuszok egymással megfeleltethetők, azonos sorrendben követik egymást. Ezt a jelenséget géntérkép és kromoszóma ko-linearitásának nevezzük. A gímszarvas géntérkép 621 marker pontja 34 kapcsoltsági csoportba rendeződik és 2532 cM hosszúságot ad (*Slate és mtsai.*, 2002a, 2. *ábra*).



2. ábra

Gímszarvas 1. kapcsoltsági csoportja. *Slate és mtsai.*, 2002a. alapján

Figure 2. Red deer Linkage Group 1., based on Slate et al., 2002a.

A fejlécben a szarvasmarha, a juh és az emberi megfelelő ortológ kromoszóma (számjelzés szerinti) olvasható le, továbbá a linkage group/kapcsoltsági csoport cM hossz értéke. A baloldalon lévő számok az egyes markerek közötti cM-ban megadott távolságok. A különböző színek a markerek típusait különítik el (kék-AFLP, piros-mikroszatellita, zöld-EST, fekete-RFLV/gén, barna-protein). Az alsó boxban feltüntetett markerekről csak azt tudni, hogy szomszédosak, vagyis nem lehet megmondani a pontos pozíciójukat a géntérképen, csupán az egyik szomszédos markertől való távolságukat adták meg.

The header contains the appropriate orthologous chromosome (numerical) for cattle, sheep and human, and the cM length value for the linkage group (blue AFLP, red microsatellite, green EST, black RFLV/gene, brown protein). The markers in the bottom box only know that they are adjacent, that is, their exact position on the gene map is not known, only their distance from one of the neighboring markers is given.

A gímszarvas genom összeállítása folyamán a szekvencia tartalmuk alapján rangsoroltuk a scaffoldokat. A kromoszómába rendezéshez először a kapcsoltsági géntérkép markerpontjait illesztettük a *de novo* scaffoldokhoz, így kaptuk meg a DNS/marker szekvenciát tartalmazó mapmarker vagyis térképpont" scaffoldokat (MMSc). Ezáltal a géntérképi pontok kiterjedt (scaffoldnyi) szekvencia környezetbe kerültek. A következő lépésben a gímszarvas térkép pon-

tok megfelelőit (ortológjait) azonosítottuk a szarvasmarha teljes genom szekvenciában. Azaz az összehasonlító géntérképezési elvet használtunk a gímszarvas adatok felől szarvasmarha megfelelők felé. A géntérképre „feltűzött" gímszarvas DNS szekvenciákat (DNS markereket, térképpontokat) illeszteni lehetett a szarvasmarha teljes genom szekvenciájára. A gímszarvas géntérkép pontjai hosszú szakaszokon azonos sorrendben, ko-lineárisan helyezkedtek el a szarvasmarha genomban is, azaz kiterjedt szinténiákat, lokális kapcsoltságokat tapasztaltunk. A következő lépésben az összehasonlító géntérképezési elvet alkalmaztunk fordított irányban: a szarvasmarha felől a gímszarvas felé. A gímszarvas DNS szekvencia scaffoldokat a szarvasmarha genom azon térképközeihez, szegmenseihez illesztettük, amelyeket az ortológ gímszarvas és szarvasmarha marker szekvenciák (azaz térképpontok) egyaránt meghatároztak, mivel a gímszarvas referencia térképi pontok szekvenciája csaknem megegyezik a két fajban, továbbá ezen pontok sorrendje azonos a két fajban, valamint a térközök/térkép szegmensek hosszai is arányosak, valamint az egyes ortológ gének szekvenciái csaknem azonosak, ezért ezeket az ortológ szarvasmarha géneket használtuk a gímszarvas scaffoldok halászatra a továbbiakban, vagyis az összehasonlító géntérképezési elvet alkalmaztunk a szarvasmarha felől-gímszarvas felé. Az ilyen módon kihalászott scaffoldokat reference gene containing scaffolds-nak (RGSc) neveztük el. Ezután felillesztettük a nem referencia géneket, rRNS, tRNS, miRNS géneket tartalmazó scaffoldokat vagyis az úgynevezett inter reference genes scaffolds-okat (IRGSc). Majd legvégül a megmaradt helyeket feltöltöttük a kimaradt 1999 bp feletti scaffoldokkal (gap filling scaffolds). Minden egyes lépésnél a szarvasmarha genomra illesztett gímszarvas scaffoldokat azonos sorrendben áttöltöttük a gímszarvas géntérkép megfelelő térképi szegmenseibe. Hosszabb DNS szekvenciák illesztésénél a következő programokat ajánlatos használni: MegaBLAST, BWA (MEM), MUMmer, NUCmer, LASTZ.

Centromeron pozíciók meghatározása

Az eukarióta sejt sejtmagjában található DNS kromoszómákba rendeződik a sejt osztódási ciklusában. A kromoszómákon megfigyelhető elsődleges befűződés a centromeron helye, amely egy rövidebb (p kar) és egy hosszabb (q kar) részre osztja a kromoszómát. A centromeronhoz tapadnak az osztódási orsó húzófonalai, amelyek szétválasztják egymástól a testvér kromatidákat, ill. a kromoszómákat a mitózis illetve a meiózis anafázisa során. Hiánya összeegyeztethetetlen a sejt életével. A centromeron helyzete alapján különböző

morfológiájú kromoszómák jönnek létre. Amikor a centromeron a kromoszóma középen található metacentrikus (M kromoszóma) kromoszómáról beszélünk. Amennyiben a kromoszóma valamelyik végéhez esik közel a centromeron a kromoszóma akrocentrikus (A kromoszóma). Az extrém végállású centromeronok helyzetét telocentrikusnak (T kromoszóma) nevezzük. A különböző kromoszóma-sávfestési eljárások beszámolnak az aktívan működő kromoszóma részekről és információval szolgálnak az egyes gének helyzetéről. A Giemsa festéssel például a lazább szerkezetű kromatinon (DNS és fehérje komplex) a folyamatosan átíródó gének régiói nem, vagy gyengén festődnek (R sávok). A tömörebb kromoszómális részek, amelyekre nem, vagy csak korlátozott mértékben jellemző a gén expresszió erősen festődő, sötét sávokat adnak a Giemsa festéssel. A centromeronok kompakt szerkezetűek, s emiatt a sötét szegmensként jelennek meg a kromoszómákon. A DNS szekvenciák orientálását nagyban elősegítik a kromoszóma festési eljárásokkal előállított citológiai képek, amelyek megmutatják, hogy akrocentrikus vagy metacentrikus kromoszómáról van-e szó.

A gímszarvas esetében készült ugyan citológiai kép a kromoszómákról, de a centromeronoknak A DNS szekvenciájához és a gének helyéhez való viszonya ez idáig feltáratlan volt. Amiben biztosak voltunk a kariogramok alapján, azaz, hogy a szarvasmarha és a gímszarvas -kromoszómák csaknem mindegyike akrocentrikus. A gímszarvas evolúciós vonalán azonban 7-9 millió évvel ezelőtt kialakult a 2n=68 kariotípusú ős, amely két akrocentrikus kromoszóma fúziójának (un. robertsoniális transzlokáció) köszönhetően egy metacentrikus kromoszómapárt hozott létre az akrocentrikusak mellett a diploid sejtben. A gímszarvasban ez a metacentrikus 5. kromoszóma. A szarvasok Y kromoszómája ősükhöz hasonlóan szubmetacentrikus maradt (*Fontana és Rubini*, 1990). A gímszarvas esetében tehát fogódzkodót jelentett a szarvasmarha és a szarvas úgynevezett „sávosra festett" kromoszómáiról készült mikroszkópi képek összehasonlítása, vagyis a gímszarvas versus szarvasmarha sávozott kromoszómák megfeleltetése. Szerencsés körülmény, hogy a gímszarvas metacentrikus (5.) kromoszóma két karja megfeleltethető a szarvasmarha két akrocentrikus (17., 19.) kromoszómájával. A szikaszarvas (amely a gímszarvas egy alfajának is tekinthető) és a szarvasmarha Giemsa festéssel nyert sávmintázatát vizsgáló, citogenetikai komparatív elemzések (*Bonnet és mtsai.*, 2001) alapján a két faj teljes kromoszómális homológiát mutat. A szarvasmarhánál korábbi kromoszóma citológiai vizsgálatok jelezték centromeron pozicióját. A későbbiekben festéssel és in situ DNS hibridizációk alapján meghatározták a centromeronok és a hozzájuk közel eső gének helyzetét (*Ma és*

mtsai., 1996). E géneket már azonosítani tudtuk a szarvasmarha genom szekvenciában, ortológjaikat pedig gímszarvas genom szekvenciában, CerEla1.0-ban. A két faj ortológ génjeinek szinténiái és a citogenetikai felvételek (a sorba rendezett metafázisos kromoszómák sávmintázatai, (*Bonnet és mtsai.*, 2001) komparatív elemzésével sikerült meghatározni a centromeronok lehetséges helyzetét a géntérképi pontokhoz igazítva mind a 34 gímszarvas kapcsoltsági csoportban. A genetikai térképek segítségével készített pszeudo-kromoszómákon be tudtuk határolni a centromeronok helyét. A centromeronok segítségével megállapítottuk, hogy 6 szarvasmarha kromoszóma kettéhasításával megfeleltethető 12 gímszarvas kromoszóma. Közülök az egyik az akrocentrikus 19. gímszarvas kromoszóma, az 1. szarvasmarha kromoszóma disztális felével feleltethető meg, míg a proximális szakasszal a 31. gímszarvas kromoszóma. Tovább bonyolítja a helyzetet, hogy 19. szarvas kromoszóma ősében egy törés és egy transzlokáció (az alsó és a felső szegmens helyet cserélt) is lejátszódott az evolúció során. Feltártunk még egy paracentrikus inverziót is a 28. gímszarvas kromoszómában és két olyan esetet, amikor egy gímszarvas kromoszóma két szarvasmarha kromoszóma tandem illlesztésével magyarázható, azaz minthatha az akrocentrikus gimszarvas kromoszómát kettétörnénk két akrocentrikus szarvasmarha kromoszómává. Az 5. gímszarvas kromoszóma metacentrikusságát citológiai és genomikai bizonyítékok egyaránt igazolták (*Bonnet és mtsai.*, 2001). A 15. gímszarvas akrocentrikus kromoszáma pedig a 28. és a 26. akrocentrikus szarvasmarha kromoszóma fúziójának (un. Robertsoniális fúzió) feleltethető meg (*Bonnet és mtsai.*, 2001).

A GENOM PROGRAMOK JELENTŐSÉGE AZ ÁLLATTARTÁSBAN ÉS AZ ÁLLATTENYÉSZTÉSBEN

Az élőlények teljes genetikai információjának megismerése sokoldalú hasznosítási lehetőséget rejt magában. A genom projektek által megismerhetővé válnak az egyes fajok evolúciós, régészeti és populációgenetikai viszonyai (*Frank és mtsai.*, 2017). Az I típusú (magáért a kitüntetett tulajdonságért felelős gén) és a II típusú (a génhez köthető, azzal együtt öröklődő speciális DNS szakasz) markerek fejlesztését nagyban megkönnyítik a rendelkezésünkre álló teljes genom szekvenciák.

A modern vadgazdálkodásban fontos szerepet tölt be az egyedazonosítás, az apai és anyai vonalak nyomon követése, amely például a szarvasfélék esetében mikroszatellitákkal (*Zsolnai és mtsai.*, 2009, *Szabolcsi és mtsai.*, 2014), illetve mitokondriális markerekkel történik meg. A 10 gímszarvas autoszomális tetranukleotid mikroszatellita lokuszra épülő „parentage

controll kit", a DeerPlex azonosító ereje 1 a 30 trillióban (10^{-16} nagyságrend) (*Szabolcsi és mtsai.*, 2014). A DeerPlex alkalmazása lehetővé teszi a forenzikus, bűnügyi, vadorzási, régészeti, természetvédelmi és vadgazdálkodási felhasználást is. A gazdaságilag fontos, mennyiségi és minőségi tulajdonságaikat meghatározó gének, genetikai struktúrák azonosítása is egyszerűbbé válik az elérhető teljes genomoknak köszönhetően. A DNS szintű markerek jól hasznosíthatók származás igazolására, ellenőrzése (tenyészbikák, versenylovak, kutyatenyésztés).

Az állati termékek előállításánál a környezeti és takarmányozási tényezőkön kívül nagy szerepet kaphatnak a minőséget és mennyiséget befolyásoló, örökíthető, poligénes vagy csupán egyetlen génnel összefüggő genetikai adottságok. Az egyetlen gén által meghatározott tulajdonságok esetében nincs szükség a teljes genom szekvencia ismeretére a markerfejlesztéshez. Ilyenkor elegendő az adott lokuszhoz tartozó strukturális eltérések feltárása. A juhok ovulációs rátáját a 6. kromoszómán elhelyezkedő *FecB* vagy *bmpr-1b* (booroola) egyetlen gén pontmutációja (a kódoló 746-os pozícióban bekövetkezett adenin-guanin csere) befolyásolja, amely szuperovulációt, ezáltal nagyobb utódszámot eredményez (*Souza és mtsai.*, 2001). A hústermelő-képesség kapcsán ismert a belga-kék és a piemonti szarvasmarhák culard jellegét (duplafarúság, túlizmoltság) okozó myostatin (MSTN) gén mutációja, amelyet a gén 3. exonjában bekövetkező 11 bázis kiesés (deléció és frameshift), vagy a guanin-adenin illetve cisztein-tirozin cseréje (pontmutáció) idéz elő (*Kambadur és mtsai.*, 1997).

Az egygénes tulajdonságokkal ellentétben a poligénes, kvalitatív és kvantitatív jellegek esetében nagyon előnyös lehet a teljes genom szekvencia vizsgálata. A tej minőségét több komponens határozza meg. A tejfehérje összetételt, a tejmennyiséget és a jól örökíthetőséget mutató tejzsírszázalékot többek között olyan nagyhatású gének befolyásolják, mint például a béta-laktoglobulint, kappa-kazeint kódoló gének, *dgat1, opn,* és *abcg2*. Bioinformatikai módszerekkel lehetőség adódik ezen génekhez kapcsolódó egypontos nukleotid polimorfizmusok (SNP) feltárására és ezaltál irányított tenyésztéssel a tejhozam és minőség fokozására (*Raschia és mtsai.*, 2018).

A gödöllői NAIK-Mezőgazdasági Biotechnológia Kutatóintézetben is folynak genomikai jellegű kutatások, amelyek új-generációs szekvenálási adatokat, teljes genom szekvenciákat használnak fel, illetve hoznak létre *de novo*. Itt hasonlították össze az őshonos mangalica fajták teljes genomját lefedő read szekvenciákat más sertés genomokkal. Ily módon sikerült a mangalica fajtajellegeket meghatározó specifikus szerkezeti különbségeket és SNP-ket azonosí-

tani. Az SNP adatbázisok nagyban segíthetik a mangalica hústermékek kvalitatív és kvantitatív PCR módszerekkel történő eredetigazolását (*Molnár és mtsai.*, 2014). Szintén Gödöllőn zajlanak a magyarországi mézelő méh populáció (pannon méh) (*Apis mellifera carnica pannonica)* teljes genomját feltáró vizsgálatok, amelyeknek célja a varroa szenzitív higiénikus viselkedésért (VSH) felelős QTL régiókhoz kötött mikroszatellita markerek kifejlesztése. A *Varroa destructor* atka által terjesztett méhvírusok a lárvákat támadják meg. A VSH-t mutató méhek eltávolítják ezeket az atkával fertőzött lárvákat a lépsejtekből, ezáltal megelőzik a vírus tovább terjedését. A krajnai méh populáció csupán 2%-ban fordul elő ez a fajta viselkedés (*Spötter és mtsai.*, 2016). A VSH-t meghatározó genetikai markerek detektálása és fejlesztése hozzájárulhat a VSH-t mutató populációk fenntartásához és szaporításához.

Az 1998-ban induló gímszarvas genom projekt a gödöllői Mezőgazdasági Biotechnológiai Központ a Kaposvári Egyetem Állattenyésztési, a Bőszenfai Szarvasfarm, az ELTE Genetikai, és a SOTE 1. Belgy. Klinika MSc és PhD programjai összefogásával jött létre. A programtervet Orosz László dolgozta ki, az állattenyésztési és vadgazdálkodási hasznosítást Horn Péter, a klinikai irányt Lakatos Péter jegyezte. Ez Magyarország első emlős genom programja, amelynek keretében elkészült a világ első kromoszómákba rendezett szarvas referencia genomja, amely online is elérhető (*Bana és mtsai.*, 2018; valamint az NCBI genom adattárában a MKHE00000000 azonosítószámmal megadva, *Link*). A gímszarvas teljes genom szekvencia ismerete számtalan lehetőséget rejt magában. Egy ezek közül a kapitális agancs genetikai hátterének feltárása, ami nem csupán vadgazdálkodási és vadászati szempontból lehet érdekes, hanem orvosbiológiai nézőpontból is. A gímszarvas agancs egyedülálló szerv, amely az éves agancs ciklus során lehullik, és újra épül, ezzel az élővilág legnagyobb mértékű csontnövekedését produkálja, miközben a gímszarvas bika fiziológiás oszteoporózist szenved el. A vázcsont ásványi anyag vesztését a nyári dőhérség alatt pótolja vissza az állat úgy, hogy őszre, a bőgés idejére jó kondícióba kerül, és tekintélyes agancsot rak fel, amely a váz csontozat 25-30%-val megegyező csonttömeget jelent. A csontfejlődésben szerepet játszó gének promóter régióinak vizsgálata során azt találta kutatócsoportunk, hogy bizonyos gének (például *col1A1*) 1 és 5 kilobázispáros promóter régiójában több runx2 transzkripciós faktor kötőhely található a gímszarvasban, mint az emberben vagy a szarvasmarhában, s talán emiatt is, ezek a gének aktívabban működnek a gímszarvasban (*Stéger és mtsai.*, 2010). A teljes genom szekvenciák nagy előnye, hogy az élőlény összes genetikai információját helyes sorrendben tartalmazzák, ezáltal megkönnyítik a laboratóriumi munkát, a kutatás-fejlesztést és nem utolsó sorban az állattenyésztést. Az online elérhető haszonállat

referencia genom adatbázisokból megtudhatjuk az adott genomra vonatkozó legfontosabb információkat és le is tölthetjük azokat

1. táblázat

Néhány fontosabb haszonállat teljes referencia genom adatai. (NCBI alapján; *Link*)

Név	Fajta	Nem	Kromoszóma szám (n)	Szekvenálási technika	Legfrissebb verzió	Genom hossz	Fehérje kódoló
Házi macska (*Felis catus*)	Abesszin	♀	18+X+MT	PacBio; 454 Titanium; Illumina; Sanger	Felis_catus_9.0 (2017)	2,52	19748
Kutya (*Canis lupus familiaris*)	Boxer	♀	38+X+MT	Sanger	CanFam3.1 (2011)	2,41	20039
Szarvasmarha (*Bos taurus*)	Hereford	♀	29+X+MT	PacBio; Illumina NextSeq 500/HiSeq/Gall	ARS-UCD1.2 (2018)	2,72	21039
Juh (*Ovis aries*)	Rambouillet	♀	26+X+MT	HiSeq X Ten; PacBio RS II	Oar_rambouillet_v1.0 (2017)	2,87	21160
Ló (*Equus caballus*)	Telivér	♀	31+X+MT	Sanger; Illumina HiSeq; PacBio	EquCab3.0 (2018)	2,51	21129
Sertés (*Sus scrofa*)	Duroc	♀	18+X/Y+MT	PacBio	Sscrofa11.1 (2017)	2,5	2079
Nyúl (*Oryctolagus*	Thorbecke	♀	21+X+MT	ABI	OryCun2.0 (2009)	2,74	20547
Baromfi (*Gallus gallus*)	Bankivatyúk (UCD001)	♀	33+W/Z+MT	Pacific Biosciences RSII	GRCg6a (2018)	1,07	17477
Ponty (*Cyprinus*	-	-	50+MT	-	common carp genome (2014)	1,71	49579
Méh (*Apis mellifera*)	(DH4)	♂	16+MT	PacBio; 10X Chromium; Bionano	Amel_HAv3.1 (2018)	0,23	9935

Table 1. Complete reference genome data for some major livestock. based on NCBI; Link

KÖVETKEZTETÉSEK

A genomika és a bioinformatika nagy hatást gyakorol az orvostudományra, a mezőgazdaságra és a környezetvédelemre egyaránt. Az élőlények teljes genetikai információjának a megismerése a genom szekvenálási módszereknek köszönhetően lehetségessé válik. A haszonállatok teljes referencia genomjának összeállításhoz szükséges módszertani ismeretek megkönnyítik a célnak megfelelő bioinformatikai programok kiválasztását és paraméterezését. A már régóta rendelkezésre álló és az újonnan összeállított, annotált referencia genomokban szintén bioinformatikai programok felhasználásával lehetőség nyílik az egyedi vagy a fajtára jellemző genetikai struktúrák azonosítására, a mennyiségi és minőségi tulajdonságokat meghatározó lokuszok megkeresésére, és

a különböző betegségekkel vagy éppen a velük szembeni rezisztenciával kapcsolatba hozható genomi részek megtalálására. Az ilyen módon megismert DNS markerek szekvenciáira primereket, azonosító kiteket lehet fejleszteni a laboratóriumi munka során. A markerekre kifejlesztett kiteket alkalmazhatják a bűnügyben, a természetvédelemben, a vadgazdálkodásban, az állattenyésztésben és az élelmiszeriparban is.

KÖSZÖNETNYILVÁNÍTÁS

A publikáció elkészítését az EFOP-3.6.3-VEKOP-16-2017-00005 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

IRODALOM

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Letöltve: Link

Bana, N.Á., Nyíri, A., Nagy, J., Frank, K., Nagy, T., Stéger, V., Schiller, M., Lakatos, P., Sugár, L., Horn, P., Barta, E., Orosz, L. (2018). The red deer Cervus elaphus genome CerEla1.0: sequencing, annotation, genes, and chromosomes. Mol. Genet. Genom., 293(3), 665-684. DOI: *10.1007/s00438-017-1412-3*

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456, 53–59. DOI: *10.1038/nature07517*

Bonnet, A., Thévenon, S., Claro, F., Gautier, M., Hayes, H. (2001). Cytogenetic comparison between Vietnamese sika deer and cattle: R-banded karyotypes and FISH mapping. Chromosome Res 9, 673-687. DOI: *10.1023/a:1012908508488*

de la Bastide, M., McCombie, W.R. (2007). Assembling genomic DNA sequences with PHRAP. Curr Protoc Bioinformatics., 17(1), 11.4.1-11.4.15. DOI: *10.1002/0471250953.bi1104s17*

Dovichi, N.J., Zhang, J. (2000). How Capillary Electrophoresis Sequenced the Human Genome. Angew Chem Int Ed Engl., 39(24), 4463-4468. DOI: *10.1002/1521-3773(20001215)39:24<4463::aid-anie4463>3.0.co;2-8*

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M, *et al.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269(5223), 496-512. DOI: *10.1126/science.7542800*

Fontana, F., Rubini, M. (1990). Chromosomal evolution in Cervidae. Biosystems, 24(2), 157-174. DOI: *10.1016/0303-2647(90)90008-o*

Frank, K., Bleier, N., Tóth, B, Sugár, L, Horn, P, Barta, E, Orosz, L, Stéger, V. (2017). The presence of Balkan and Iberian red deer (Cervus elaphus) mitochondrial DNA lineages in the Carpathian Basin. Mammal Biol., 86, 48-55. DOI: *10.1016/j.mambio.2017.04.005*

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., Jaffe, D.B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci., 108(4), 1513-1518. DOI: *10.1073/pnas.1017351108*

Gyurján, I., Molnár, A., Borsy, A., Stéger, V., Hackler, L., Zomborszky, Z., Papp, P., Duda, E., Deák, F., Lakatos, P., Puskás, L.G., Orosz, L. (2007). Gene expression dynamics in deer antler: mesenchymal

differentiation toward chondrogenesis. Mol Genet Genomics, 277(3), 221–235. DOI: 10.1007/s00438-006-0190-0

Kambadur, R., Sharma, M., Smith, T.P., Bass, J.J (1997). Mutations in myostatin (GDF8) in double-muscled Belgian Blue and Piedmontese cattle. Genome Res., 7(9), 910-915. DOI: 10.1101/gr.7.9.910

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M., Goldstein, M.M., Grigoriev, I.V., Hackett, K.J., Haussler, D., Jarvis, E.D., Johnson, W.E., Patrinos, A., Richards, S., Castilla-Rubio, J.C., van Sluys, M.A., Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences of the United States of America, 115(17), 4325–4333. DOI: 10.1073/pnas.1720115115

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Link

Ma, R.Z., Beever, J.E., Da, Y., Green, C.A., Russ, I., Park, C., Heyen, D.W., Everts, R.E., Fisher, S.R., Overton, K.M., Teale, A.J., Kemp, S.J., Hines, H.C., Guérin, G., Lewin, H.A. (1996). A male linkage map of the cattle (Bos taurus) genome. J Hered., 87(4), 261-271. DOI: 10.1093/oxfordjournals.jhered.a022999

Maxam, A.M., Gilbert, W. (1977). A new method for sequencing DNA. Proc Natl Acad Sci., 74(2), 560-4. DOI: 10.1073/pnas.74.2.560

Molnár, J., Nagy, T., Stéger, V., Tóth, G., Marincs, F., Barta, E. (2014). Genome sequencing and analysis of Mangalica, a fatty local pig of Hungary. BMC Genomics, 15(1), Article: 761. 1-12. DOI: 10.1186/1471-2164-15-761

Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P., Barron, A.E. (2011). Landscape of next-generation sequencing technologies. Anal. Chem., 83(12), 4327–4341. DOI: 10.1021/ac2010857

Pevsner, J. (2015). Bioinformatics and functional genomics. John Wiley & sons inc., UK. pp 1124.

Raschia, M. A., Nani, J. P., Maizon, D. O., Beribe, M. J., Amadio, A. F., és Poli, M. A. (2018). Single nucleotide polymorphisms in candidate genes associated with milk yield in Argentinean Holstein and Holstein x Jersey cows. Journal of Animal Science and Technology, 60(1), Article: 31. 1-10. DOI: 10.1186/s40781-018-0189-1

Rusk, N. (2011). Torrents of sequence. Nat. Methods, 8(1), 44. DOI: 10.1038/nmeth.f.330

Sanger, F., Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol., 94(3), 441-448. DOI: 10.1016/0022-2836(75)90213-2

Sanger, F., Nicklen, S., Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A., 74(12): 5463–5467. DOI: 10.1073/pnas.74.12.5463

Shulaev, V., Sargent, D., Crowhurst, R. *et al.* (2011). The genome of woodland strawberry (Fragaria vesca). Nat Genet., 43(2), 109-116. DOI: 10.1038/ng.740

Slate, J., Van Stijn, T.C., Anderson, R.M., McEwan, K.M., Maqbool, N.J., Mathias, H.C., Bixley, M.J., Stevens, D.R., Molenaar, A.J., Beever, J.E., Galloway, S.M., Tate, M.L. (2002). A deer (subfamily Cervinae) genetic linkage map and the evolution of ruminant genomes. Genetics, 160(4), 1587-1597.

Souza, C.J., MacDougall, C., MacDougall, C., Campbell, B.K., McNeilly, A.S., Baird, D.T (2001). The Booroola (FecB) phenotype is associated with a mutation in the bone morphogenetic receptor type 1 B (BMPR1B) gene. J Endocrinol., 169(2), R1-6. DOI: 10.1677/joe.0.169r001

Spötter, A., Gupta, P., Mayer, M., Reinsch, N., Bienefeld, K (2016). Genome-wide association study of a Varroa-specific defense behavior in honeybees (Apis mellifera). Journal of Heredity, 107(3), 220-227. DOI: 10.1093/jhered/esw005

Stéger, V., Molnár, A., Borsy, A., Gyurján, I., Szabolcsi, Z., Dancs, G., Molnár, J., Papp, P., Nagy, J., Puskás, L., Barta, E., Zomborszky, Z., Horn, P., Podani, J., Semsey, S., Lakatos, P., Orosz, L. (2010). Antler development and coupled osteoporosis in the skeleton of red deer Cervus elaphus: expression dynamics for regulatory and effector genes. Mol Genet Genomics, 284(4), 273-287. DOI: 10.1007/s00438-010-0565-0

Stein, L. (2001). Genome annotation: from sequence to biology. Nat. Rev. Genet., 2(7), 493–503. DOI: *10.1038/35080529*

Sutton, G.G., White, O., Adams, M.D., and Kerlavage, A.R. (1995). TIGR Assembler: A new tool for assembling large shotgun sequencing projects. Genome Sci. Technol., 1(1), 9-19. DOI: *10.1089/gst.1995.1.9*

Szabolcsi, Z., Egyed, B., Zenke, P., Padar, Z., Borsy, A., Steger, V., Pasztor, E., Csanyi, S., Buzas, Z., Orosz, L., (2014). Constructing STR multiplexes for individual identification of Hungarian red deer. J. Forensic Sci., 59(4), 1090-1099. DOI: *10.1111/1556-4029.12403*

Wang, B., Ekblom, R., Bunikis, I., Siitari, H., Höglund, J. (2014). Whole genome sequencing of the black grouse (Tetrao tetrix): reference guided assembly suggests faster-Z and MHC evolution. BMC genomics, 15(1), Article: 180. 1-13. DOI: *10.1186/1471-2164-15-180*

Zsolnai, A., Lehoczky, I., Gyurmán, A., Nagy, J., Sugár, L., Anton, I., Horn, P., Magyary, I. (2009). Development of eight-plex microsatellite PCR for parentage control in deer. Arch. Tierz., 52(2), 143-149. DOI: *10.5194/aab-52-143-2009*